ISSUES IN THE SAFETY OF COMPLEX SYSTEMS

NAGW-1333

By:

V.J. Ree, Jr.
L.K. Lauderbaugh

Department of Electrical, Computer and Systems Engineering
Department of Mechanical Engineering, Aeronautical Engineering & Mechanics
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

CIRSSE Document #10

Issues in the Safety of Complex Systems

Vincent J. Ree Jr. *
and
L. Ken Lauderbaugh **

* Student Member IEEE
Graduate Student, Department of Mechanical Engineering, Aeronautical Engineering, and Mechanics, Rensselaer Polytechnic Institute, Troy, NY 12180-3590

** Assistant Professor, Department of Mechanical Engineering, Aeronautical Engineering, and Mechanics, Rensselaer Polytechnic Institute, Troy, NY 12180-3590

Abstract

In this paper, an architecture for an automated complex system is described which involves splitting up subsystem components into an operational element and a safety monitor. This is done in order to optimize the safety and efficiency of the system. The safety process (detection, evaluation, resolution) is described and classification schemes for threats to safety are proposed. These clasification threat schemes are used in the detection and evaluation aspects of that process. Briefly, classification is along three lines: necessary response speed (immediateness of threat), risk from threat, and location of hazards and victims of threats. Examples are given which justify these concepts. Ideas for further research are given.

1. Introduction

Automated systems are becoming increasingly vital and complex. In fact, many tasks require automated systems, such as robots, because the environments involved are exceedingly hazardous for human beings. An example of such a complex system is that of the Flight Telerobotic Servicer (FTS) under development at NASA for use on the Space Station. [1] The FTS is to have very versatile capabilities and be able to evolve over time as new needs arise. It will be able to perform many different tasks including satellite retrieval and aid in the construction of the space station among other tasks to be determined as the system evolves. However, such a complex system must be both safe and reliable during its operational lifetime.

During the operation of such systems, situations may arise which may be considered hazardous. The importance of how such a situation is resolved is dependent upon many factors including: the seriousness of the threat (Risk), speed of response, capabilities in responding, side-effects, etc. If a system faces a situation that has a large risk factor, it is inherently more

vital that the optimal decisions for resolution are made. Even if the correct decisions are made, real-time implementation issues are just as important.

Such similar situations and questions are also commonly addressed by human beings. Unfortunately, humans do not always make the best decisions. Reasons why humans may make suboptimal decisions include: lack of or overload of information, lack of expertise, oversight, emotional influence, etc. For these reasons, humans are perhaps not the ideal sources for such decisions, especially if the situation is an emergency; the stress involved coupled with all the factors associated with a complex system can easily overwhelm any individual. [2]

Therefore, the best way to handle hazardous situations involving complex systems would be to automate the 'safety process' simply because decisions involving the resolution of such situations can be made more objectively, swiftly, and completely (that is to say all available relevant information is used). Furthermore, the response can be planned and executed

much faster than by a human.

By the 'safety process', what is meant is this: searching for possible hazardous situations, evaluating such situations, planning and executing the response to avoid or defuse them (resolution), and then returning to normal. The safety process is one which readily involves the principles of artificial intelligence and expert systems.

Ramirez [3,4] has discussed a system for safe intelligent robotic control. This system contained two parts: (1) An executive controller (EC) which "planned and commanded the activities of the robot and other machines for the accomplishment of a goal." [4] , and (2) a Robotic Intelligent Safety System (RISS), which actively interacts with the EC and "monitors the activities in the environment for failure detection and abnormal system behaviors. The RISS diagnoses the failures and it generates corrective actions." [4] Ramirez' work primarily concerned safety in terms of collision avoidance of a single robot with obstacles in the environment using "stratified risk"

and "forbidden volumes". [3,5,6] The concepts developed however are, in principle, extendable to multiple safety objectives and robots (i.e. a complex system).

Dodhiawala et al recognized that many systems, such as the Space Station, are actually composed of many different interactive and interdependent subsystems. Therefore, they concluded that a "safety advisor" must interact with these subsystems to take appropriate action after diagnosing a potentially hazardous situation. They also concluded that the subsystems should be integrated using a blackboard-based architecture which "...appears appropriate to the task of effectively organizing the activity and behavior of diverse subsystems." [7] How this "safety advisor" performs and is implemented is an important point we address.

In this paper, we discuss some of the implementational issues involved in the intelligent control of safety of complex systems. We begin by defining and justifying a set of classification schemes for the threats associated with a complex system or any of its subsystems and demonstrate with examples.

The ideas presented are very generic and extendable to any definition of system, automated or human. We then discuss an architecture to optimize the resolution of safety threats. Questions for further research are then proposed.

2. Classes of Threats

Before we continue with the classification of threats to a complex system, we need to make some clarifying definitions:

(1) A HAZARD is a source of damage. A situation is hazardous or DANGEROUS if a hazard is threatening.

(2) A THREAT is an expression of a hazard to do damage.

(3) DAMAGE or HARM is the result of an executed threat. If a system is damaged, it functions in a manner not considered desirable. A 'system' may be taken in the most abstract sense if desired.

(4) A VICTIM is a system that receives damage.

(5) The RISK associated with a threat may be loosely

defined as the product of the damage and the probability of that damage occuring in a given period of time. It is a measure of the seriousness of a threat.

(6) A system is SAFE if there is a low probability that it will be damaged.

With these definitions in place, we can now classify threats along three different lines:

(1) Immediateness of threats

(2) Risk and solvability associated with threats

(3) Locations of hazards and of victims

2.1 Purpose for Classification

These classifications types are made for reasons related to the 'safety process' as performed by an automated complex system. The safety process involved detection, evaluation, and resolution of threats to safety of the system and its environment. The first two classification schemes are useful for evaluation and

resolution. The third classification scheme is useful in dividing up the process of detecting threats.

2.2 Immediateness of Threats

A threat may be classified according to the response time neccessary for its resoultion. An immediate threat is one which the time to resolve the threat is considerable when compared to the time in which the threat will probably cause damage. Usually such threats are unexpected and require a "reflex response". Non-immediate threats usually are expected, but even if they are not, the time involved in resolution is such that the response is planned.

Threats caused by hazards may be both unexpected and expected. Expected threats can usually be dealt with over time through planning and gradual resolution. Such threats, while they may be emergencies, can be dealt with most effectively. These are non-immediate threats. An example of a situation involving an expected threat would be that involving obstacle avoidance for a robot.

Unexpected threats may come in two forms. Typically, many possible threats and hazards are expected and thus plans of action may already be in place whether the threat exists or not. However, unexpected threats might exist which are, in a sense, 'perturbations' (or worse) from the 'standard threats'. Therefore, the plans of action would have to be similarly adjusted which involves extra time to plan and implement resolutions. If such time is not available, a reflex response may be necessary; that is the threat is immediate. If the time is available, then the solution is planned; the threat is non-immediate. A reflex response may be suboptimal using many criteria, but given a time constraint, proper planning and implementation may be impossible. Ramirez noted these facts in the development of RISS; it contained a reflex module which was separated from the "inference engine" which, among other tasks, was responsible for planning. [4] An example of a sudden unexpected threat is that of "robot runaway".

2.3 Risk and Solvability

Risk, as stated above, is a measure of the seriousness of a threat. It is necessary to assign risks associated with threats so that an automated safety system can determine the order in which the threats should be resolved (Threat Scheduling). Risk assessment should not be delegated with threats that require reflex responses however. Such threats should be addressed first since response time is essential for the resolution of those threats.

Solvability is another essential issue in terms of threat scheduling. There may be many ways to resolve a threat. The optimum resolution is that which yields no damage and requires the least time. However, there may not exist a solution which can satisfy these criteria. Therefore, it seems necessary to seek a solution that minimizes the damage and implementation time. Solvability is a measure of the effort necessary for a particular solution.

In terms of threat scheduling, what would be desired is to

minimize the overall projected damage from all threats. For example, a threat with high risk but also low solvability should be scheduled after a threat with somewhat less risk but a higher solvability. As a heuristic to threat scheduling, we can define an 'adjusted risk' roughly as the product of risk and solvability, with those threats having the highest adjusted risk being scheduled earlier. These ideas are basically abstract extensions of those pertaining to triage from surgical medicine.

2.4 Locations of Hazards and Victims

Sato and Inoue discussed a method of hazard identification for human-robot systems using action-changes and action-chains models. They identified six classes of action-types in the action-changes model which can yield a classification scheme for threat types. [8] We propose that, from a system point-of-view, there are three types of threats based upon the locations of the hazards and of the victims:

(SE) Hazard: System Victim: Environment

(ES) Hazard: Environment Victim: System

(SS) Hazard: System Victim: System

2.4.1 Example of a Type SE Threat:

Consider a robot in operation and a human enters the workspace. Subsequently, the human is now in a dangerous situation since the robot might harm the human for a variety of reasons. The robot must now determine what it can do so as not to harm the human. Schematically, a system (robot) is threatening a potential victim (human) that is in the environment (workspace).

2.4.2 Example of a Type ES Threat:

Consider an airplane on autopilot flying along a designated flight path. However, the flight path will intersect a severe storm. The question here is how can the flight path be altered to avoid the storm ahead. Schematically, the system (airplane) is the potential victim since it is being threatened by a hazard (the storm) in its environment.

2.4.3 Example of a Type SS Threat:

Type SS threats involve the system damaging itself. Intuitively, it would seem that such threats could be decomposed into Type SE or Type ES threats among the subsystems of that system. However, when a system is broken down far enough, analysis along such lines is fruitless, but instead may be approached as a question of reliability. If a system is currently operating in a satisfactory manner, but is unreliable, then it will soon be damaged. In short, the system is threating itself by operating and if it continues to do so, it may become a hazard to its environment setting up further Type SE, Type SS, or even Type ES threats that might not have existed otherwise.

For instance, consider a professional athlete such as a football player, say a running back. Suppose further that he has sustained a minor injury such as a pulled muscle. In terms of performance, this player may not be as considered reliable as he

would be had he been healthy. If he continues to play, he has a higher probability of committing an error, such as a fumble. Even worse, he might sustain a more serious injury which would require him to be sidelined for a longer period of time than if he had not played in order to heal himself. Here, the athlete, (along with the coach, team physician, etc.) must make the decision as to whether or not he should play. The athlete represents the system in this example. He is able to 'monitor' himself, and thus estimate his own reliability. Automated systems should have similar capabilities so that they can judge for themselves if they are damaged and thus unsafe.

It is interesting to note that, in the football player example, the decision to play is also influenced by other factors such as the importance of the current game and the health of the other players. Extending these ideas to complex systems, the question arises as to what is the acceptable system reliability for operation in given a situation. If a situation is critical, it may be necessary to have the system be operational even if the reliability is less than that in a normal situation.

Subsequently, safety may be compromised. Again, these ideas are abstractions of triage.

3. Aspects of Safety Implementation

As stated earlier, Ramirez [4] discussed a system for safe robotic control which was composed of an executive controller and an intellegent safety system for a single robot. However, complex systems are typically divided into subsystems to perform different tasks. These subsystems are often divided into further subsystems to perform the required subtasks. The question arises as to how should safety be 'controlled' in such a complex system.

In answering this question there are two guiding criteria:

(1) The response to a safety problem should be as swift as possible.

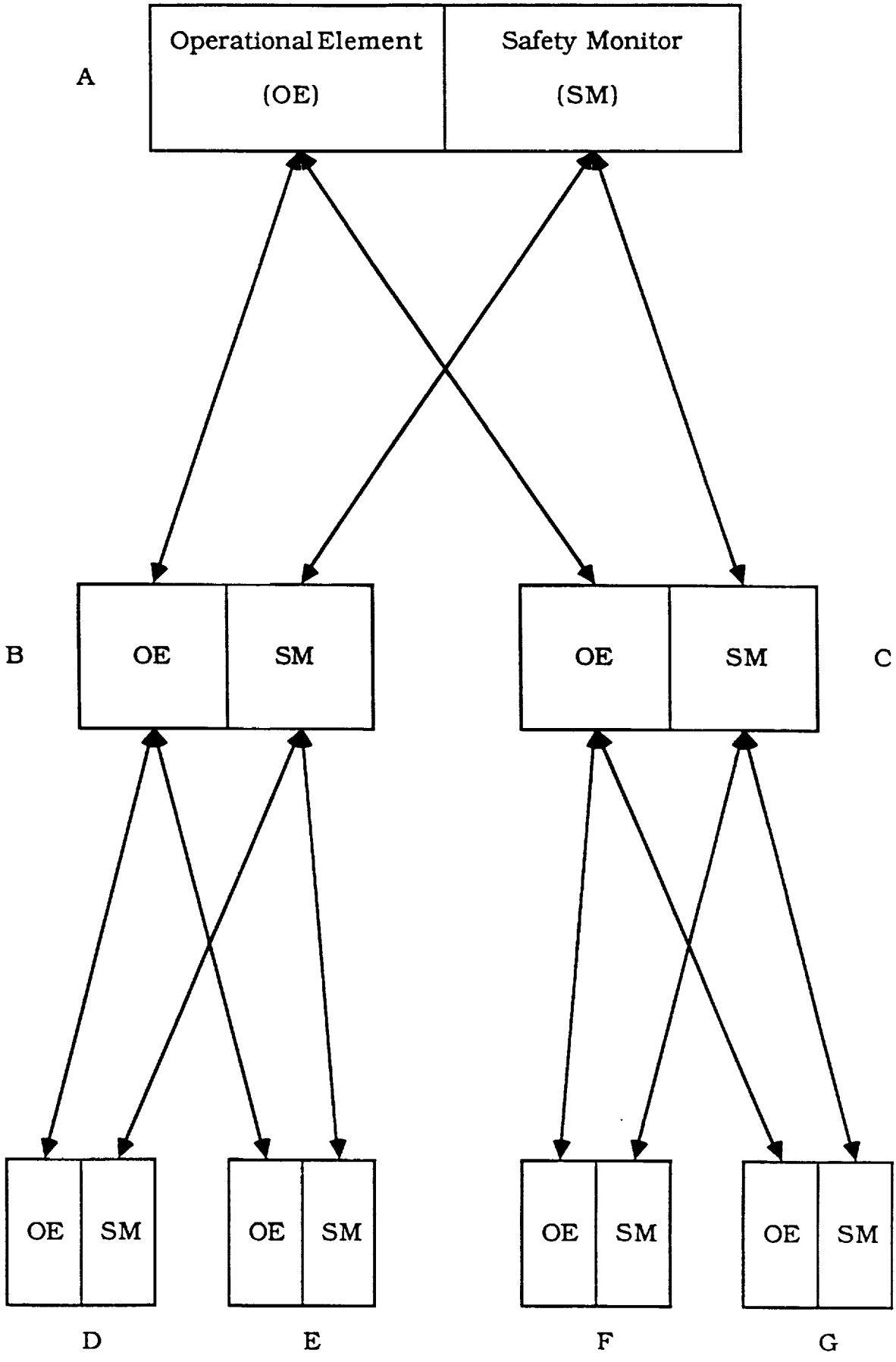
(2) Safety problems should be contained so as to keep them from getting worse and to keep overall system efficiency as high as possible, that is avoid interfering with other systems so that they can still

perform.

For these two reasons, it seems logical to monitor safety at every level. This implies that every subsystem be composed of two elements: (1) an operational element that actually performs any designated tasks, and (2) a subsystem safety monitor which checks the safety with respect to that subsystem and is informed of the safety of any of that subsystem's subsystems.

Figure 1 represents a system (A) which contains two subsystems (B and C) to perform its required task. This implies that the task involved requires subtasks performed by these subsystems. These are further broken down to another layer of subtasks and subsystems (D,E and F,G). Each subsystem has an operational element and a safety monitor. This is comparable the the scheme described in [4]. Each safety monitor checks that subsystem's safety and then informs the superior system of the status of that subsystem. If the subsystem being monitored has a immediate threat, then its safety monitor executes a reflex response immediately. However, if the threat is less immediate,

Figure 1.


Operational Element
(OE)
Safety Monitor
(SM)
A
B
OE
SM
OE
SM
C
OE
SM
OE
SM
OE
SM
OE
SM
D
E
F
G

the safety monitor consults with the superior system's safety monitor for information and planning. If need be, the process continues to an even higher safety monitor, etc. This compartmentalizes the problem and optimizes the response time by having communication occur only between the necessary systems. It is evident that, if each subsystem is to have its own safety monitor and operational element, the overall system structure is that of a networked blackboard system [7,9].

Using the FTS as an example, consider the retrieval of a satellite. Assume, for simplicity, that only four sybsystems are involved in this task: Navigation/Propulsion, Power, Sensing, Grasping. Let's assume that in the power subsystem, there are two redundant battery units. Suppose one is in use, and the other is not and that the first one is expected to die soon. Once the safety monitor (a simple sensor here) detects this threat (Type SS, which could chain to further Type SE threats; non-immediate; low-risk and high solvability), it alerts the safety monitor for the power system which switches batteries. If also this battery was not fully charged, the power system safety

monitor would alert the FTS safety monitor so it can take action to conserve energy such as degrading the performance of the sensor system until it was vital, etc. This threat has much higher risk and is also non-immediate; planning is involved for the best resolution.

4. Conclusion

In this paper, we discussed some issues involved in implementing an intelligent safety system from the system's perspective. Classification schemes for threats involving the system were proposed which may be used to aid the system in detecting, evaluating, and resolving these threats. To optimize safety and efficiency, every subsystem should be split into an operational element and a safety monitor for that element. Responses to threats are either reflexive or planned. Planning is done in consultation with other relevant safety monitors. This structure is implicitly a network of automated blackboard safety expert systems. The ideas presented are applicable to any

generic system.

Mentioned, but not fully discussed, were some ideas which should be developed through further research:

(1) Risk, Solvability, and Adjusted Risk.

The purpose of the ideas behind these terms was to aid an intelligent safety system order which threats to safety it should address (Threat Scheduling). These concepts should be more quantifiable.

(2) Acceptable Operational Reliabilty.

As stated in section 2.4, the acceptable operational reliability in a given situation is dependent upon the importance of the situation, as well as the threats, involved. There is a need to quantify the dependence of acceptable reliability on these factors.

(3) Reliability Analysis Versus Systen Decomposition

In section 2.4.3, it was stated there exists a point where reliability analysis is better than system

decomposition in terms of analyzing safety. Research involving determining these points is important for economic and hardware reasons.

References

[1] Flight Telerobotic Servicer (FTS) Strawman Concept Engineering Report, NASA Goddard Space Flight Center (GSFC), SS-GSFC-0031, March 15, 1987.

[2] Critical Issues in Robot-Human Operations During the Early Phases of the Space Station Program, Lauderbaugh, L.K.; Montgomery, T. Davetta; Kondraske, George V.; Hoard, Katy; Walker, Michael W.; Kim, Dong-Min; Chang, Kai-Hsiung; Cross, James; Dannelly, Steve; Consortium for Space/Terrestrial Automation and Robotics, February 26, 1988.

[3] "Robot Intelligent Safety System", Ramirez, Carlos A., Proceedings of the 14th ISIR/Robots 8 Conference, Vol. 2, pp. 19-50 to 19-62, 1984.

[4] "Artificial Intelligence Applied to Robot Fail-Safe Operations", Ramirez, Carlos A., Proceedings of RI/SME Robots 9, Vol. 2, pp. 19-21 to 19-37, 1985.

[5] "Stratified Levels of Risk for Collision-Free Robot Guidance", Ramirez, Carlos A., Proceedings of the 15th

ISIR Conference, pp. 959-966, 1985.

[6] "Robotic Obstacle Avoidance Using a Camera and a 3-D Laser Scanner", Bennekers, Robert, and Ramirez, Carlos, Proceedings of the Robots 10 Conference, pp. 3-51 to 3-63, 1986.

[7] "Integrating Architecture for Complex System Design", Dodhiawala, R.T., Jagannathan, V., Baum, L.S., Robotics and Expert Systems-1986: Proceedings of ROBEXS '86, pp. 75-80, 1986.

[8] "Safety Assessment of Human-Robot Systems (1st Report, Hazard Identification Based on the Action-Changes and Action-Chains Models)", Sato, Y., and Inoue, K., Bulletin of JSME, Vol. 29, No. 256, pp. 3618-3625, October 1986.

[9] "Blackboard Systems: The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures", Nii, H. Penny, AI Magazine, pp. 38-53, Summer 1986.